Claims

What is claimed is:

1. A method comprising the steps of:

providing a set of sequences, wherein the sequences are not aligned;

discovering a plurality of patterns common to a plurality of the sequences; and

determining if a candidate sequence comprises a predetermined number of the patterns.

2. The method of claim 1, wherein the patterns common to a plurality of the set of sequences comprise test patterns, wherein the sequences in set of sequences comprise test sequences, and wherein the step of determining if a candidate sequence comprises a predetermined number of the patterns comprises the step of determining if there are candidate patterns in the candidate sequence that match all of the predetermined number of test patterns.

3. The method of claim 1, further comprising the step of determining if each of the plurality of patterns is statistically significant.

4. The method of claim 1, wherein the step of discovering is performed without any knowledge about properties or features of sequences in the set of unaligned sequences.

5. The method of claim 1, further comprising the steps of if the candidate sequence comprises the predetermined number of patterns, adding the candidate sequence to the set of sequences to create a new set of sequences and performing the step of discovering on the new set of sequences.

6. The method of claim 1, wherein each sequence comprises a series of symbols and wherein each pattern comprises a plurality of positions, some of the positions each comprising at least one expected symbol and other of the positions comprising "don't care" positions.

7. The method of claim 6, wherein, for one of the positions, the at least one expected symbol is a plurality of expected symbols.

8. The method of claim 3, wherein the step of determining if each of the plurality of patterns is statistically significant comprises the steps of selecting one of the patterns, determining if a probability that the selected pattern occurs in a sequence meets a predetermined threshold, and continuing to select additional patterns until each pattern has been selected.

9. The method of claim 8, wherein the step of determining if a probability that the selected pattern occurs in a sequence meets a predetermined threshold further comprises the steps of using a second-order Markov chain method to determine the probability that the selected pattern occurs in a sequence and determining a natural logarithm of the probability that the selected pattern occurs in a sequence.

10. The method of claim 3, wherein the step of determining if each of the plurality of patterns is statistically significant further comprises the steps of removing instances of each of the patterns from the set of sequences to create a new set of sequences and performing the step of discovering on the new set of sequences.

11. The method of claim 3, wherein the step of determining if each of the plurality of patterns is statistically significant further comprises the steps of if any of the patterns is statistically significant, selecting a statistically significant pattern, modifying a composite descriptor to include the selected pattern if the selected pattern is not already part of the composite descriptor, and continuing to select statistically significant patterns until all statistically significant patterns have been selected.

12. The method of claim 1, wherein the step of discovering a plurality of patterns common to a plurality of the sequences comprises the steps of:

selecting a predetermined threshold that indicates how many of the sequences should contain a pattern for the pattern to be considered common;

discovering patterns, if any, that are common to the predetermined threshold of sequences;

if there are no patterns common to the predetermined threshold of sequences, decreasing the predetermined threshold; and

performing, until the predetermined threshold is less than a predetermined amount, the step of discovering patterns, if any, that are common to the predetermined threshold of sequences and the step of if there are no patterns common to the predetermined threshold of sequences, decreasing the predetermined threshold.

13. A method for unsupervised building and exploitation of composite descriptors, the method comprising the steps of:

i. providing a training set of sequences, each sequence comprising a plurality of symbols;

ii. determining a set of maximal patterns, each of the maximal patterns being common to a predetermined number of the sequences, wherein the step of determining a set of maximal patterns is performed without any knowledge about properties or features of sequences in the set of unaligned sequences;

iii. determining which, if any, of the maximal patterns are statistically significant; and

iv. creating a composite descriptor from the statistically significant maximal patterns.

14. The method of claim 13, wherein the sequences in the training set are unaligned.

15. The method of claim 13, wherein the step of creating a composite descriptor from the statistically significant maximal patterns further comprises the steps determining which of the statistically significant maximal patterns are currently not part of the composite descriptor, adding those statistically significant maximal patterns that are currently not part of the composite descriptor to the composite descriptor, and removing the added statistically significant maximal patterns from the training set of sequences.

16. The method of claim 15, wherein each symbol comes from an alphabet that describes DNA (deoxyribonucleic acid) or proteins.

17. The method of claim 13, wherein the symbols are numerical.

18. The method of claim 15, further comprising the steps of iterating steps (ii) through (iv) until either the training set contains no sequences or there are no statistically significant maximal patterns common to the sequences in the training set.

19. The method of claim 15, further comprises the step of determining if a candidate sequence comprises a predetermined number of the statistically significant maximal patterns.

20. The method of claim 19, comprising the steps of if the candidate sequence comprises the predetermined number of the statistically significant maximal patterns, adding the candidate sequence to the set of sequences to create a new training set of sequences and performing the steps (ii) through (iv) on the new training set of sequences.

21. The method of claim 13, wherein the step of determining which, if any, of the maximal patterns are statistically significant comprises the step of determining for each of the maximal patterns if a probability that this maximal pattern occurs in a sequence meets a predetermined threshold.

22. The method of claim 13, wherein the set of maximal patterns is empty and wherein the step of determining a set of maximal patterns further comprises the steps of reducing the predetermined number of sequences and performing step (ii) again.

23. A system comprising:

a memory that stores computer-readable code; and

a processor operatively coupled to said memory, said processor configured to implement said computer-readable code, said computer-readable code configured to:

provide a set of sequences, wherein the sequences are not aligned;

discover a plurality of patterns common to a plurality of the sequences; and

determine if a candidate sequence comprises a predetermined number of the patterns.

24. A system for unsupervised building and exploitation of composite descriptors, comprising:

a memory that stores computer-readable code; and

a processor operatively coupled to said memory, said processor configured to implement said computer-readable code, said computer-readable code configured to:

i. provide a training set of sequences, each sequence comprising a plurality of alphabetic symbols;

ii. determine a set of maximal patterns, each of the maximal patterns being common to a predetermined number of the sequences, wherein the maximal patterns are determined without any knowledge about properties or features of sequences in the set of unaligned sequences;

iii. determine which, if any, of the maximal patterns are statistically significant; and

iv. create a composite descriptor from the statistically significant maximal patterns.

25. An article of manufacture comprising:

a computer readable medium having computer readable code means embodied thereon, said computer readable program code means comprising:

a step to provide a set of sequences, wherein the sequences are not aligned;

a step to discover a plurality of patterns common to a plurality of the sequences; and

a step to determine if a candidate sequence comprises a predetermined number of the patterns.

26. An article of manufacture for unsupervised building and exploitation of composite descriptors, comprising:

a computer readable medium having computer readable code means embodied thereon, said computer readable program code means comprising:

a step to provide a training set of sequences, each sequence comprising a plurality of alphabetic symbols;

a step to determine a set of maximal patterns, each of the maximal patterns being common to a predetermined number of the sequences, wherein the maximal patterns are determined without any knowledge about properties or features of sequences in the set of unaligned sequences;

a step to determine which, if any, of the maximal patterns are statistically significant; and

a step to create a composite descriptor from the statistically significant maximal patterns.